Cornier Alexandre
04/11/2019

université de BORDEAUX

Jukes Cantor 1969 Model

Introduction :

Phylogenetics is the study of evolutionary relationships between organisms. The main purpose is to reconstruct correct phylogenetics tree while estimating the time of divergence between organisms and knowing the sequence of events along evolutionary lineages.

So, to build phylogenetics tree it's necessary to calculate the genetic distance between two homologous sequences. Which is the number of substitutions accumulated between them since they diverged from their common ancestor. The Challenge is not simply count the number of position at which the two sequence differ but to calculate the genetic distance while taking into account the fact that there may be several substitutions on the same site

For this, many probabilistic models vary according to the parameters used to describe the velocities at which one nucleotide replaces another during evolution. As part of my project, I worked on the Jukes Cantor 1969 model to show how this model of substitution is the simplest and fastest one as it provides results in one single pass. Also this model specifies that the equilibrium frequencies of the four nucleotides are 25% each and has the same probability to be replaces by any other during the evolution.

2019

Cornier Alexandre
04/11/2019

Plan :

1. Material & Methods
2. Results
3. Discussion & Conclusion
4. References

1. Material & Methods :

As for this project, it's all about simulating the genetic distances, the only material needed is a computer with a development language to build and run the models.
The programming language chosen in this case is Python3, because of its simplicity and powerful functions embedded in complementary modules. I've used the Python3 environment provided by the Anaconda distribution version xx, in particular:

- Python 3.7.0 as the command interpreter
- Spyder 3.3.6 as Integrated Development Editor
- Modules
    - Random
    - Math
    - MathPlotLib

Only Python and Emacs where used as the beginning but I moved to Anaconda to simplify the installation of the modules and to improve the development speed by leveraging the Spyder IDE.
Regarding the methods, I've proceeded in multiple steps to build the functions required to create simulated sequences, to compute the genetic distances, and then to compare with the Jukes Cantor 1969 model.

Step 1:

To create DNA sequences, I've used a set of functions that takes randomly letters from a nucleic alphabet for the length specified as parameter.
Another function creates mutated sequences successively, from the previously created one as initial sequence. This list can be used as Data source for the next experimentations.

Step 2:

I can now create different sets of mutated sequences, with different length and number of substitutions.
Based on this, I've developed the code that provide the Hamming Distances between sequences, for a given number of experiments.
Those results will be used to compare the behaviour of the Jukes Cantor model with the roughly computed Data.

Step 3:

Having the material ready to experiment and to compare with, I've coded the Jukes Cantor 1969 model represented by the formula below:

$$T = -\frac{3}{4}\ln\left(1 - \frac{4}{3}d\right)$$

- ln = logarithm
- d = sites that differ between the two sequences the Jukes-Cantor estimate of the evolutionary distance

The JC69 model makes several assumptions:

- The probability of changing from one state to a different state is always equal,
- The different sites are independent.

Only the mutation rate is taken in account for computing the Genetic Distances.
Compared to other Genetic Distances models, it provides fast results in one single pass.

The Function I've codded accordingly to the JC69 model use the following arguments:

- means: mean of hamming distance for instance, passed as integer
- L: sequence length

The last part of this Project was to represent graphically the Hamming Distances and Jukes Cantor 1969 Distances vs. a given number of substitutions. I've used, still in Python3, the MathPlotLib module to perform several experiments with different parameters to observe the behaviour of the JC69 model.

2. Results :

Because the sequences are created randomly, the results may vary between each run. However, the variation is not really significant when using the same parameters.

Here is an example run with the following parameters:

- Sequence Length: 1000
- Number of Experiments: 10
- Range of number of Substitutions: from 100 to 2200 step 200

I've used a range of substitutions to represent graphically the Genetic Distances computed as Hamming Distances and JC69 Distances.

In the graph below, the blue curve represents the Hamming Distance as a logarithm, and I observe that the red curve representing the JC69 model is closed to a mathematical right. I will discuss about this behaviour later on, in this document.

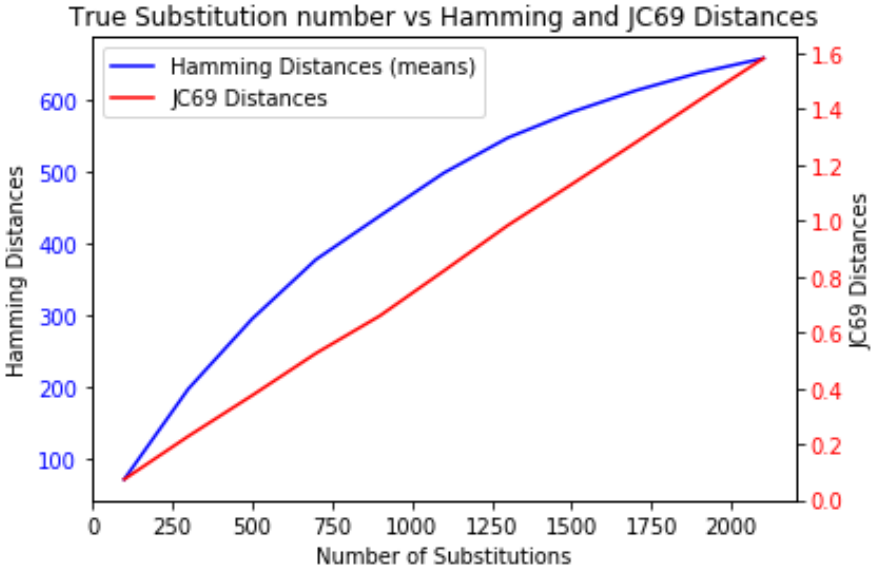

Figure 1: Graphic represent the true substitution number vs Hamming and JC69 Distances

This graph can be used to estimate the distances for a number of Substitutions.
It can also provide an estimated count of Substitutions for a given Distance, that may be useful as the actual number of substitutions in a sequence is often underestimated (figure 1)

3. Discussion & Conclusion :

To go further in analysis of the behaviour of the Jukes Cantor 1969 model, compared to the Hamming Distances, I've ran several times the program with different parameters.

Run 1

Using parameters:

- Sequence Length: 1000
- Number of Experiments: 10
- Range of number of Substitutions: from 100 to 5000 step 200

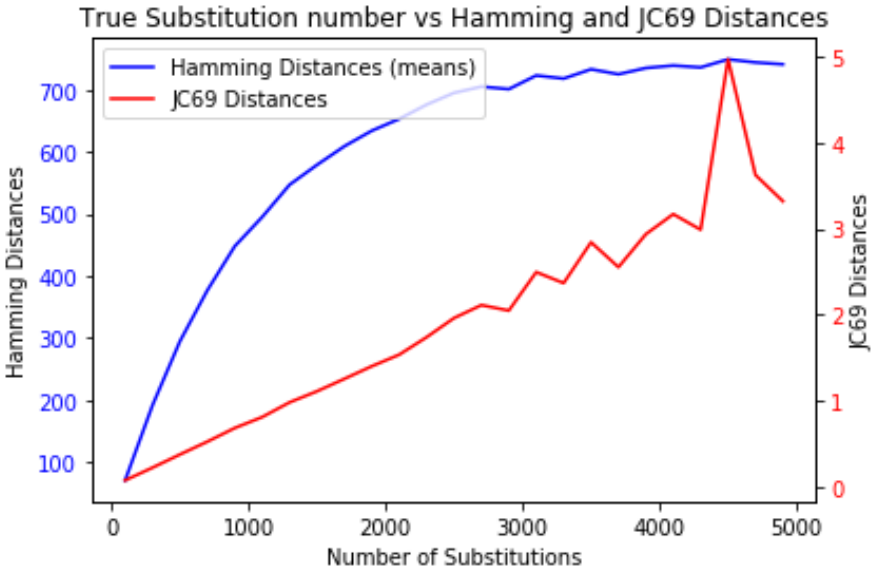

Figure 2 : Graphic represent the true substitution from 100 to 5000 number vs Hamming and JC69 Distances

For this run, I observe an erratic misalignment in the linear relationship occurring at the end of the red curve (figure 2). It happens when the Hamming Distances curve reaches a tray, just above a count of Substitutions corresponding to 70% of the Sequence length (700 / 1000).

It happens when multiple substitutions per sites have been accumulated. Those sequences are called “Saturated” and this is where the precise shape of the curve depends on the details of the substitution model used.

Run 2

Using parameters:

- Sequence Length: 5000
- Number of Experiments: 10
- Range of number of Substitutions: from 100 to 5000 step 200

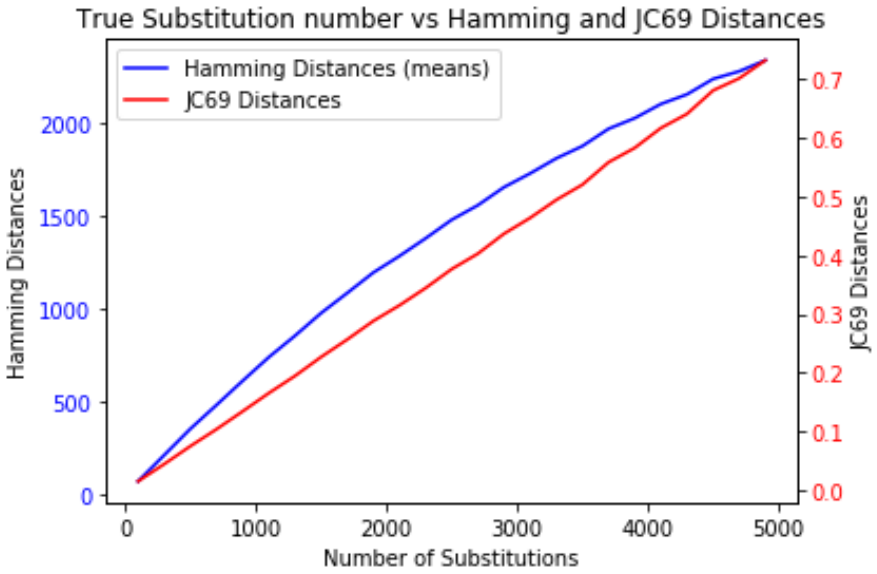

Figure 3 : Graphic represent the true substitution number vs Hamming and JC69 Distances with a sequence length of 5000

Increasing the Sequence length, while keeping the range for the number of Substitutions, create a graph similar to the initial one, providing a balance that preserves the accuracy of the JC69 model (figure 3).

Run 3

Using parameters:

- Sequence Length: 5000
- Number of Experiments: 50
- Range of number of Substitutions: from 100 to 5000 step 200

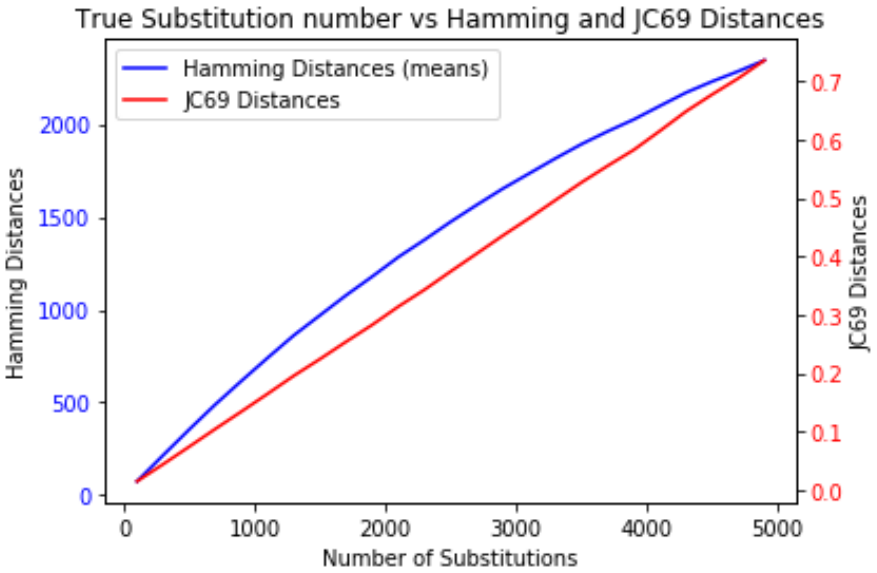

Figure 4 : Graphic represent the true substitution number vs Hamming and JC69 Distances with a sequence length of 5000 and 50 experiments

This run was done to estimate the impact of the number of Experiments on the curves. There are no significant changes in values, only the curves are less noisy (figure 4).

Conclusion

The Jukes Cantor 1969 model has a couple of advantages:

- The Data needed for the simulation is simple to build
- It's fast to compute the result
- The accuracy is good since the number of substitutions doesn't create saturated sequences. The sequences of equal length must match for more than ¼ of their sites

However, there's some ways of improvements as the only parameter of this model is the overall substitution rate, and it's not sufficient for the calculation of the evolutionary distance under the more complex models.

It's the first substitution model offering a genetic distance correction. Many others have been developed in the 1980s and 1990s.

The Jukes Cantor 1969 model is still very interesting to simulate a genetic evolution in a fast way. It also provides a good reference to challenge other methods or algorithms.

4. References :

Cavalli-Sforza, L., Edwards, A.W.F. (1967), “Phylogenetic analysis. Models and estimation procedures”, Evolution, 21, 550-570.

Felsenstein, J. (1981), “Evolutionary trees from DNA sequences: a maximum likelihood approach”, Journal of Molecular Evolution, 17, 368-376.

Aydin, B., Pataki, G., Wang, H., Bullitt, E., & Marron, J. S. (2009). A principal component analysis for trees. The Annals of Applied Statistics, 3, 1597– 1615.

Benny Chor, Michael D. Hendy, Sagi Snir, Molecular Biology and Evolution, Volume 23, Issue 3, March 2006, Pages 626–632

Christopher Tuffley, W. Timothy, J. White, Michael D. Hendy, David Penny, Molecular Biology and Evolution, Volume 29, Issue 12, December 2012, Pages 3703–3709